Bespoke Instruments: A New Tool for Addressing Unmeasured Confounders

David B. Richardson and Eric J. Tchetgen Tchetgen

Presenter: Will Zhang

University of Washington

May 1, 2023

Table of contents

1. Model and Setting
2. Proposed Method
3. Sufficient bespoke IV identification conditions
4. Simulation

Motivation

Interests: quantifying an exposure-disease causal association in a setting where the exposure, disease, and some potential confounders of the association of interest have been measured.

Problem: Epidemiology studies often miss information on potential confounders.

Solution: use a reference population without treatment exposure and repurpose the measured confounders as a bespoke instrument to account for unmeasured variables.

Model of interest

Define variables:

- A is a point exposure or treatment.
- Z denote measured variables that confound the A-Y association.
- U denote unmeasured variables that confound the A-Y association.
- U need not be associated with Z; however, we allow that U and Z may be associated.

Potential interested estimate:

- average treatment effect among the treated, $E\left[Y^{a=1} - Y^{a=0} \mid A = 1\right]$
- Under $Y^a \perp A \mid U, Z$, if (U, Z) and standard positivity assumption, we can find ATT or ATE.

DAG

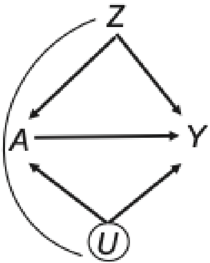

Figure 1. Illustration of the relationship between exposure, outcome, measured confounder, and unmeasured confounder. The undirected arrow represents the possible presence of a causal relationship between *Z* and *U* or of an unmeasured common cause of *Z* and *U*.

Reference Population and Prognostic Score

Require a “hypothetical external intervention” a priori:

- Let $R = 1$ denote the reference population, lacking treatment exposure opportunity.

  EX: Calendar time, Spatial, and physical reasons preventing exposure.
- $R = 0$ denotes the target population.
- Random sampling from super-population is not required, but we require key characteristic invariant across sub-populations.

Prognostic score $F(Z)$ denotes the expected value of Y conditional on Z among reference group, $E[Y \mid Z, R = 1] = F(Z) = E\left[Y^{a=0} \mid Z\,,\ R = 1\right]$, by treatment exclusion.

Bespoke instrumental variable

In study (target) population:

- compute $\hat{F}(Z)$ and $\tilde{Y} = Y - \hat{F}(Z)$ for all individuals
- Under conditions, we expect $\hat{F}(Z)$ captures $E\left[Y^{a=0} \mid Z, R=0\right]$, in which case $\tilde{Y}$ and Z can be expected to be mean independent in the study population.
- Above induce instrument-like property for control of U.

Requirement:

- requires that the $Z - Y$ association in the reference population matches the $Z - Y^{a=0}$ association in the target population.
- weaker assumption than fully exchangeable factors between reference and target population (measured and unmeasured).

Additive model: Two-stage regression

First, we obtain the predicted value of A given Z,

$$\hat{A}(Z) = \hat{E}(A \mid Z).$$

Second, fit a regression of Y on $\hat{A}(Z)$ including $\hat{F}(Z)$ as an offset,

$$\mathrm{E}(Y \mid \hat{A}(Z), \hat{F}(Z)) = \beta_0 + \beta_1^{IV} \hat{A}(Z) + \hat{F}(Z)$$

where

- $\hat{\beta}_1^{IV}$ is the estimate of interest
- $\beta_0 = E\left[Y^{a=0} \mid R = 0,\ Z = 0\right] - E\left[Y^{a=0} \mid R = 1, Z = 0\right]$ accounts for the possibility that the baseline mean for the treatment-free potential outcome might differ between target and reference populations.

Multiplicative model: Two-stage regression

For binary Y, use 2-stage linear-logistic regression:

- estimate the difference in the log of the counterfactual outcome mean $\log E\left(Y^a \mid Z\right)$ per unit change in a
- Under assumptions, estimates do not suffer from confounding by U.

First, we obtain the predicted value of A given Z,

$$\hat{A}(Z) = \hat{E}(A \mid Z).$$

Second, we fit a logistic regression of Y on $\hat{A}(Z)$ including logit $[\hat{F}(Z)] = \text{logit}(\hat{\Pr}(Y = 1 \mid Z, R = 1))$ as an offset,

$$E(Y \mid \hat{A}(Z), \hat{F}(Z)) = \text{expit}\left(\beta_0 + \beta_1^{IV}\hat{A}(Z) + \text{logit}[\hat{F}(Z)]\right)$$

G-estimation

Additive model and multiplicative model can proceed Similarly.

1. For an additive model $E\left[Y^{a} - Y^{a=0} \mid A = a, Z\right] = \gamma_1 a$.
2. For a multiplicative model $\log\left(\frac{E[Y^{a}|A=a,Z]}{E[Y^{a=0}|A=a,Z]}\right) = \gamma_1 a$.

G-estimation of the proposed bespoke IV estimator is done by identifying parameter estimates that result in a lack of association between the instrument Z, and

1. $H(\hat{\gamma}_1, \hat{\gamma}_0) = \tilde{Y} - \hat{\gamma}_1 A - \hat{\gamma}_0$
2. $H(\hat{\gamma}_1, \hat{\gamma}_0) = \widetilde{Y}\exp(-\hat{\gamma}_1 A - \hat{\gamma}_0) - 1$, where $\widetilde{Y} = Y/\hat{F}(Z)$

Partial bespoke instrumental variable

Suppose that $Z = (Z_1, Z_2)$, under assumptions

- we take Z_1 only as a bespoke instrumental variable
- Z_2 may not be valid IV, so they become additional covariates that we adjust for.

Specifically, in the second stage of 2-stage least squares under an additive model, we fit a regression of Y on $\hat{A}\left(Z_1, Z_2\right)$ and Z_2, including $\hat{F}(Z) = \hat{F}\left(Z_1, Z_2\right)$ as an offset,

$$\begin{aligned} \mathrm{E}\left(Y \mid \hat{A}\left(Z_1, Z_2\right), \hat{F}\left(Z_1, Z_2\right), Z_2\right) = {} & \beta_0 + \beta_1^{IV} \hat{A}\left(Z_1, Z_2\right) \\ & + \hat{F}\left(Z_1, Z_2\right) + \beta_2 Z_2 \end{aligned}$$

where $\hat{\beta}_1^{IV}$ is the estimate of interest, and $\beta_0 + \beta_2 Z_2$ is a model for $E\left[Y^{a=0} \mid R = 1, Z_1 = 0, Z_2\right] - E\left[Y^{a=0} \mid R = 0, Z_1 = \ 0, Z_2\ \right]$ (note the risk/outcome mean under no exposure differs as a function of Z_2 in target and reference populations conditional on $Z_1 = 0$).

Identification conditions

1. Consistency, such that $E\left[Y^{a} \mid A=a, R=0, Z_2\right] = E\left[Y \mid A=a, R=0, Z_2\right]$ if $A=a$.
2. A degenerate reference population with $R=1$, in which we have $E[Y \mid R=1, Z] = E\left[Y^{a=0} \mid R=1, Z\right]$.
3. Partial population exchangeability, such that $E\left[Y^{a=0} \mid R=0, Z_1=1, Z_2\right] - E\left[Y^{a=0} \mid R=0, Z_1=0, Z_2\right] = E\left[Y^{(a=0)} \mid R=1, Z_1=1, Z_2\right] - E\left[Y^{(a=0)} \mid R=1, Z_1=0, Z_2\right]$.
4. Partial homogeneity (i.e., no interaction between A and Z_1) in causing the outcome among the treated, such that $E\left[Y^{a} - Y^{a=0} \mid A=a, Z_1=1, Z_2, U, R=0\right]$ $= E\left[Y^{a} - Y^{(a=0)} \mid A=a, Z_1=0, Z_2, U, R=0\right]$.
5. Bespoke instrumental variable relevance: $E[A \mid R=0, z_1, z_2\ ]$ depends on z_1 for each observed z_2.

Result 1

Under conditions 1-5, we have that $E\left[Y^{a=1} - Y^{a=0} \mid A=1, R=0, Z_2\right]$ is uniquely identified from the empirical distribution of (Y, A, Z) in the study population and (Y, Z) in the reference population. It is in fact given by the bespoke IV estimand:

$$E\left[Y^{a=1} - Y^{a=0} \mid A=1, R=0, Z_2\right] \\ = \frac{E\left[Y - F(Z) \mid Z_1 = 1, R=0, Z_2\right] - E\left[Y - F(Z) \mid Z_1 = 0, R=0, Z_2\right]}{E\left[A \mid Z_1 = 1, R=0, Z_2\right] - E\left[A \mid Z_1 = 0, R=0, Z_2\right]}.$$

Remarks

- assumption 3 is weaker than than conditional population exchangeability of the target and reference populations (i.e., $E\left[Y^{a=0} \mid R = \ 0, Z_1, Z_2\right] = E\left[Y^{a=0} \mid R = 1, Z_1, Z_2\right]\big)$, or full population exchangeability that the joint distribution of $\left(Y^{a=0}, Z_1, Z_2\right)$ is the same in both populations.
- unlike a standard IV, with "bespoke" IV, Z_1 does not require independent of U conditional on Z_2 for identification purposes of the effect of treatment on the treated given Z_2.
- condition 4 is analogous to a no-interaction assumption routinely made in the IV setting

setup

- 1,000 studies, with 2,500 people in each subpopulation.
- 4 measured covariates: $Z_1, \ldots, Z_4$, and 1 unmeasured covariate: U
- Z_1, Z_3, and U were binary, and Z_2 and Z_4 were continuous from a uniform $(-1, 1)$ distribution.
- assigned A as a random binary variable that took a value of 1 with probability $1/\left(1 + \exp\left(-\left(-0.1 - 0.5 \times Z_1 - 0.5 \times Z_2 - 0.5 \times Z_3 - 0.5 \times Z_4 + 1 \times U\right)\right)\right)$

Two Scenarios:

1. Y, was a continuous variable that took a value of $(1 + 1 \times Z_1 + 1 \times Z_2 + 1 \times Z_3 + 1 \times Z_4 + 1 \times U + 1 \times A + \varepsilon)$, where $\varepsilon \sim N(0, 1)$.
2. Y was a binary variable that took a value of 1 with probability $0.1 + 0.05 \times Z_1 + 0.05 \times Z_2 + 0.05 \times Z_3 + 0.05 \times Z_4 + 0.2 \times U + 0.2 \times A$.

Models compared

2 marginal structural regression models with stabilized inverse probability of exposure weights.

1. the first marginal structural model, the estimated propensity of exposure was derived from a logistic model fitted to each simulated cohort for A as a function of $Z_1 - Z_4$.
2. Same as 1 while including U as measured.

Results are compared by

- mean of the estimates
- estimated standard deviation of the estimates (the empirical standard error, or ESE)
- square root of the mean of squared difference between the estimated associations and the specified true effect of A on Y (the root mean squared error, or RMSE).

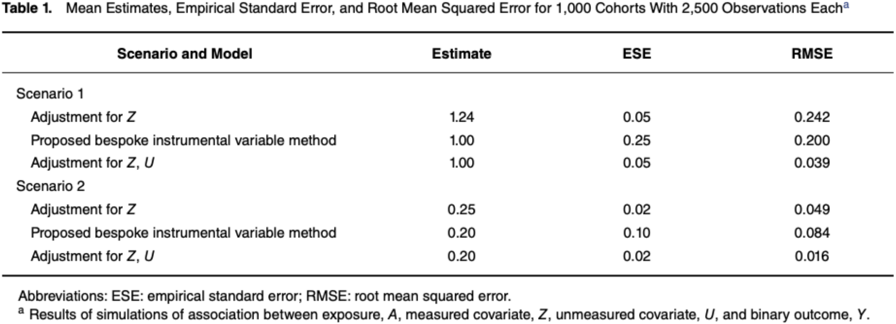

Table 1. Mean Estimates, Empirical Standard Error, and Root Mean Squared Error for 1,000 Cohorts With 2,500 Observations Each[a]

Scenario and Model	Estimate	ESE	RMSE
Scenario 1			
Adjustment for *Z*	1.24	0.05	0.242
Proposed bespoke instrumental variable method	1.00	0.25	0.200
Adjustment for *Z*, *U*	1.00	0.05	0.039
Scenario 2			
Adjustment for *Z*	0.25	0.02	0.049
Proposed bespoke instrumental variable method	0.20	0.10	0.084
Adjustment for *Z*, *U*	0.20	0.02	0.016

Abbreviations: ESE: empirical standard error; RMSE: root mean squared error.

[a] Results of simulations of association between exposure, *A*, measured covariate, *Z*, unmeasured covariate, *U*, and binary outcome, *Y*.

Empirical Study

Target: 86,611 people who were present in Hiroshima or Nagasaki at time of bombings
Reference: 26,531 people who were away from the cities at the time of the bombings
Variables Includes:

- Measure of Exposure to Radiation defined by weighted DS02 colon dose
- Age of death, city, and sex.

Empirical Example

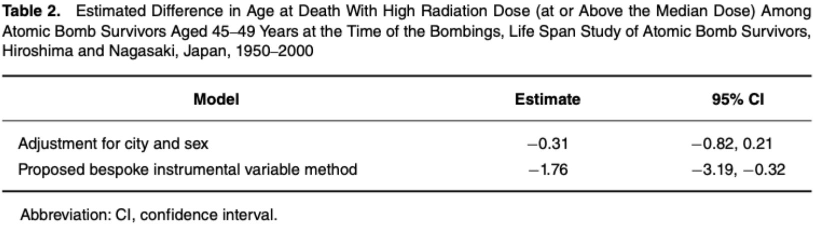

Table 2. Estimated Difference in Age at Death With High Radiation Dose (at or Above the Median Dose) Among Atomic Bomb Survivors Aged 45–49 Years at the Time of the Bombings, Life Span Study of Atomic Bomb Survivors, Hiroshima and Nagasaki, Japan, 1950–2000

Model	Estimate	95% CI
Adjustment for city and sex	−0.31	−0.82, 0.21
Proposed bespoke instrumental variable method	−1.76	−3.19, −0.32

Abbreviation: CI, confidence interval.

Thank you